View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Eldorado - Ressourcen aus und für Lehre, Studium und Forschung

On the Complexity of Rule Discovery from Distributed Data

Martin Scholz

University of Dortmund, 44221 Dortmund, Germany,
scholz@ls8.cs.uni-dortmund.de,
http://www-ai.cs.uni-dortmund.de/

September 14, 2005

Abstract

This paper analyses the complexity of rule selection for supervised learning in distributed scenarios. The selection of rules is usually guided by a utility measure such as predictive accuracy or weighted relative accuracy. Other examples are support and confidence, known from association rule mining. A common strategy to tackle rule selection from distributed data is to evaluate rules locally on each dataset. While this works well for homogeneously distributed data, this work proves limitations of this strategy if distributions are allowed to deviate. To identify those subsets for which local and global distributions deviate may be regarded as an interesting learning task of its own, explicitly taking the locality of data into account. This task can be shown to be basically as complex as discovering the globally best rules from local data. Based on the theoretical results some guidelines for algorithm design are derived.

1 Introduction

The induction of interesting rules from classified examples has been studied extensively in the Machine Learning literature throughout the last decades.

A variety of metrics like predictive accuracy, precision, or the binomial test function have been suggested to formalise the notions of interestingness and usefulness of rules. [5] gives an overview of different metrics and illustrates the differences by means of ROC isometrics. There are several learning tasks that can be formulated as optimisation problems with respect to a specific metric. Classifier induction and subgroup discovery are two examples. Usually it is assumed that all the available data is accessible to a single learner. In this case the metrics allow to identify a set of patterns that maximise the selected utility function. The amount of data necessary to identify the best rules with high probability can be considered as an indicator of complexity from an information theoretic point of view. Different sample bounds have been proven for different commonly applied metrics [9].

There are several learning scenarios with restricted access to the available data. In the domain of knowledge discovery in databases, for example, the data is often split to different sites and may not be communicated at the level of single examples. Among the reasons are privacy issues and costs.

Learning tasks can be adopted to distributed scenarios in various ways. The objective of this work is to analyse the corresponding increase in complexity, compared to non-distributed learning. To this end distributed variants of rule selection are investigated. Due to its generality the task of subgroup discovery fits nicely into this framework. It allows to specify the utility function used for pattern selection as a parameter [6]. Each subgroup is usually represented by a Horn logic rule, so utility functions are specific kinds of rule selection metrics. This paper investigates in which situations a local evaluation of rules may help to identify globally best rules, and how corresponding learning tasks are related to each other.

The remainder of this paper is organised as follows. Section 2 repeats the formal definition of non-distributed subgroup discovery. This task is extended to distributed data in section 3, assuming a homogeneous distribution at all sites. In section 4 this assumption is weakened in two ways, which are both shown to increase computational complexity to find a set of approximately best rules in the worst case. Additionally, a bound for the maximum deviation of commonly used utility functions is proved. This motivates the task of relative local subgroup discovery, which is introduced and analysed in section 5. Section 6 discusses how the presented tasks are related to distributed boosting and distributed frequent itemset mining. After discussing some practical considerations towards specific algorithmic solutions, section 7 summarises and concludes.

2 Standard subgroup discovery

This sections discusses the formal background for non-distributed supervised learning, especially for subgroup discovery. Given is a set of m classified examples $\mathcal{E} := \langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle$ from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ defines the instance space and $\mathcal{Y}$ the set of labels. The representation language ($\mathcal{H}$) used in this work contains logical rules, denoted as $A \rightarrow C$. Each antecedent A is identified with its corresponding subset of $\mathcal{X}$ (or $\mathcal{E}$, respectively), while each conclusion C consists of a label from $\mathcal{Y}$.

Formally rules are evaluated with respect to a distribution function over $\mathcal{X}$. This work confines itself to descriptive learning, so given a single database or example set $\mathcal{E}$ it is often appropriate to assume a uniform distribution D over $\mathcal{E}$.

The next definitions provide the building blocks for utility functions.

Definition 1 *The* coverage (Cov) *of a rule $A \rightarrow C$ under distribution D is defined as the probability that the rule is applicable for an example $\langle x, y \rangle$ randomly sampled $\sim D$:*

$$\text{Cov}_D(A \rightarrow C) := Pr_{\langle x,y \rangle \sim D}\left[x \in A\right]$$

Definition 2 *The* bias *of a rule $A \rightarrow C$, $C \in \mathcal{Y}$ under D is defined as the difference between the conditional probability of C given A and the default probability (class prior) of C:*

$$\text{Bias}_D(A \rightarrow C) := Pr_{\langle x,y \rangle \sim D}\left[y = C \mid x \in A\right] - Pr_{\langle x,y \rangle \sim D}\left[y = C\right]$$

These two definitions allow to state a very general class of utility functions.

Definition 3 *A function $f : \mathcal{H} \times D \rightarrow \mathbb{R}$ satisfying the following constraint for all $r, r' \in \mathcal{H}$ is called a* utility function*:*

$$\begin{aligned} &(\text{Cov}_D(r) \geq \text{Cov}_D(r')) \wedge (\text{Bias}_D(r) \geq \text{Bias}_D(r') > 0) \\ &\Rightarrow f(r, D) \geq f(r', D) \end{aligned}$$

Additionally, if one of the inequalities is strict, then $f(r, D) > f(r', D)$.

The most commonly used class of utility functions in the scope of subgroup discovery [6] is given by the following definition.

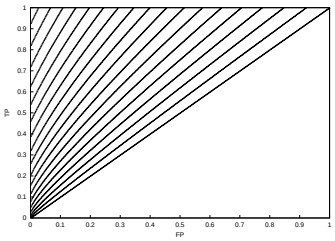

Figure 1: $\alpha = 1/2$

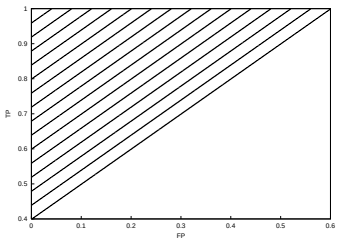

Figure 2: $\alpha = 1$

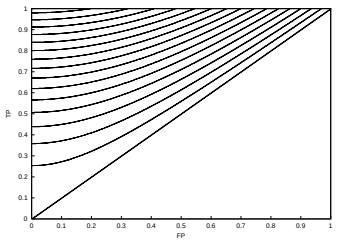

Figure 3: $\alpha = 2$

Definition 4 *For a given parameter α and distribution D the utility (or quality) $\mathrm{Q}_D^{(\alpha)}$ of a rule $r \in \mathcal{H}$ is defined as*

$$\mathrm{Q}_D^{(\alpha)}(r) := \textsc{Cov}_D(r)^\alpha \cdot \textsc{Bias}_D(r).$$

The parameter α allows for a data- and task-dependent trade-off between coverage and bias. Definition 4 covers metrics which are factor-equivalent to the binomial test function ($\alpha = 0.5$), weighted relative accuracy ($\alpha = 1$), and a function commonly used to put higher emphasis on coverage ($\alpha = 2$). Fig. 1-3 show the corresponding isometrics in ROC space [4]. Each point in the plot refers to a false positive (x-axis) and a true positive rate (y-axis), which reflect the fractions of correctly and incorrectly covered examples. A line in a ROC diagram consist of performances for which the metric yields the same score.

Definition 3 is broad enough to also cover predictive accuracy, which is equivalent to $\mathrm{Q}_D^{(1)}$ for binary prediction tasks with equal default probabilities for both classes, and which is still monotone in Cov and Bias, otherwise. The similarity between rule selection metrics for different skew ratios is discussed in [3]. In association rule mining [1] rules are filtered (or pruned) by their support (Cov) and confidence. The latter is monotone in the Bias, although the default probability is usually ignored. When support and confidence are combined (respecting monotonicity) to find a ranking of most interesting rules, then this problem can also be considered as a specific case of subgroup discovery.

For a specific choice of the utility function, the goal of subgroup discovery is to identify a set of n best or approximately best rules. One algorithm solving this problem exactly is MIDOS [10]. It works on relational data and searches the hypothesis space exhaustively, except for safe pruning.

3 Homogeneously distributed data

A first extension towards distributed subgroup discovery is to assume that several sets of data are available, which all obey a common underlying probability distribution. One can think of the different sets as generated by bootstrapping from a single, global dataset. In such a case local and global subgroups are basically identical. However, due to statistical deviations caused by bootstrapping and the smaller size of example sets, some of the rules with lower global utility might be found among the n best subgroups evaluated locally at each site.

Choosing $\mathrm{Q}^{(1)}$ (Def. 4), the probability that the utility function deviates locally from the true (global) value by more than a fixed constant $\epsilon \in \mathbb{R}^+$ can be bounded by Chernoff's inequality. This probability decreases exponentially with a growing number of examples. Sample bounds have been proven for different utility functions [9], especially for $\mathrm{Q}^{(\alpha)}$ with $\alpha \in \{.5, 1, 2\}$. As a brief summary one can state that the estimates behave well for reasonably large sample sizes, a constraint which is e.g. safely met in the context of distributed databases. Accordingly, the n-best subgroups problem has been adopted to a probabilistic scenario, in which utility functions are evaluated using i.i.d. samples [9]:

Definition 5 *Let $\delta \in (0, 1)$ denote a given minimum confidence and $\epsilon \in \mathbb{R}^+$ denote a given maximal error. Then the* approximate n-best hypotheses problem *is to identify a set G of n hypotheses from a hypothesis space $\mathcal{H}$, such that with confidence $1 - \delta$*

$$(\forall h' \in \mathcal{H} \setminus G) : \mathrm{Q}(h') \leq \min_{g \in G} (\mathrm{Q}(g) + \epsilon)$$

The results reported for this problem directly apply to homogeneously distributed datasets. For large local datasets the probability of missing a subgroup that is globally much better than the locally best ones is reasonably small.

It is worth to note, that there are also some practically relevant evaluation metrics that do not allow to tackle the approximate n-best hypotheses problem by adaptive sampling. One example is the chi-square test function, for which sampling-based utilities estimates can be arbitrarily far from the true utilities [9]. For these utility functions distributed subgroup discovery from local data is intractable. The following sections focus on functions $\mathrm{Q}_D^{(\alpha)}$.

4 Inhomogeneously distributed data

Subgroup discovery for homogeneously distributed data can be tackled and analysed using the same techniques as in the non-distributed setting. This section addresses the situation in which data is split to different sites, but no distributional assumption can be made. First of all the notation for different databases is introduced.

The example set $\mathcal{E}$ is composed of k subsets $\mathcal{E}_1, \ldots, \mathcal{E}_k$ that were sampled from different probability distributions. Let D_i denote the distribution at site i for the corresponding example set $\mathcal{E}_i \subseteq \mathcal{E}$, and let D denote the global distribution over $\mathcal{E}$. D is a weighted average of the local distributions.

The *local* Cov and Bias of a rule $A \to C$ at site i can be expressed in terms of definition 1 and 2, replacing D by D_i. For example

$$\textsc{Bias}_{D_i}(A \to C) := Pr_{\langle x,y\rangle \sim D_i}\left[y = C \mid x \in A\right] - Pr_{\langle x,y\rangle \sim D_i}\left[y = C\right]$$

refers to the local Bias at site i. Accordingly, a local utility function evaluates each rule $A \to C$ by

$$\mathrm{Q}^{(\alpha)}_{D_i}(A \to C) = \left[\textsc{Cov}_{D_i}(A \to C)\right]^{\alpha} \cdot \textsc{Bias}_{D_i}(A \to C)$$

The first task stated in this setting is to find subgroups that globally perform well, given a discovery procedure that evaluates rules locally. If for instance the globally best rule appears poor at any site, then it obviously needs to perform even better at some other. For this reason one could expect that the globally best rules are easily found at the local sites, even if the local distributions differ. A similar property eases frequent itemset mining from distributed data, because it allows for safe pruning in the case of skewed data [2].

In the case of homogeneously distributed data as discussed in section 3, the marginal distributions over $\mathcal{X}$ and the conditional probabilities of the target given $x \in \mathcal{X}$ were identical at all sites. In order to quantify by how much each of these assumptions is weakened the following definitions are useful.

Definition 6 *Two distributions $D_1, D_2 : \mathcal{X} \to \mathbb{R}^+$ are called* factor-similar up to γ *for an $A \subset \mathcal{X}$ and $\gamma > 1$, if*

$$(\forall x \in A) : \gamma^{-1} \leq \frac{D_i(x)}{D(x)} \leq \gamma.$$

Definition 7 *For an $A \subseteq \mathcal{X}$ two joint distributions $D_1, D_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ are called* conditionally similar up to ϵ, $\epsilon > 0$, *if*

$$(\forall \langle x, y \rangle \in A \times \mathcal{Y}) : \left| \frac{D_1(x,y)}{D_1(x)} - \frac{D_2(x,y)}{D_2(x)} \right| \leq \epsilon.$$

Please recall that utility functions are defined based on distributions underlying the example sets. For this reason definitions 6 and 7 do not require the same set of examples to be observable at all sites to allow for finite bounds.

The following theorem shows, that if the assumption of homogeneously distributed data made in section 3 is weakened at all, then it is possible to obtain drastically different sets of best rules when evaluating a quality function globally and locally.

Theorem 1 *Let G_i denote the set of n best rules for each site $i \in \{1, \ldots k\}$ ($k \geq 2$), given an arbitrary utility function. Let G denote the set of n best rules with respect to the global distribution. Then it is possible in the general case, that every $x \in \mathcal{X}$ is covered by at most one ruleset from $\{G, G_1, \ldots, G_k\}$, where a ruleset is said to cover x if one of its elements does. This statement even holds in the following two cases:*

1. *The global and local marginal distributions of $\mathcal{X}$ are equivalent, and global and local joint distributions of $\mathcal{X} \times \mathcal{Y}$ are conditionally similar up to an arbitrarily small $b > 0$.*

2. *For all local sites $i \in \{1, \ldots, k\}$ the conditional distributions of $\mathcal{X} \times \mathcal{Y}$ are identical, and each local marginal distribution of $\mathcal{X}$ is factor-similar to the global one up to an arbitrarily small $\gamma > 1$ for any subset of $\mathcal{X}$.*

Proof
It is sufficient to generically construct an example for both specific cases. The following proofs apply to all utility functions, but require some assumptions about the set $\mathcal{H}$ of possible hypotheses. These assumptions are met for the logical rules commonly used to characterise subgroups.

First the theorem is proved for the case of equal marginal distributions. The idea is to "prepare" for each site $i \in \{1, \ldots, k\}$ a set S_i of n disjoint subsets of $\mathcal{X}$: $S_i = \{R_{i,1}, \ldots, R_{i,n}\}$. For the global view a separate set $S_0 = \{R_{0,1}, \ldots, R_{0,n}\}$ of n rules is prepared. Let the common marginal

distribution D assign equal weight to each subset, so that all rules with antecedent $R \in \bigcup_{i=0}^{k} S_i$ have the same coverage Cov. All reasonable utility functions increase monotonically with the Bias in this case. Let C denote an arbitrarily chosen class and b and ϵ small, strictly positive real values. The joint distribution $D_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ at site i is constructed so that

$$\text{Bias}_{D_i}(R_{p,j} \to C) = \begin{cases} b/k + \epsilon, \text{ for } p = 0 \text{ (global)} \\ b \qquad\quad , \text{ for } p = i \text{ (local)} \\ 0 \qquad\quad , \text{ for } p \notin \{0, i\} \end{cases}$$

for all $1 \leq j \leq n$. The joint global distribution $D : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is computed as the average of the joint local distributions, since the marginal distributions are assumed to be equivalent. Hence the Bias of every "local rule" $R_{i,j} \to C$, $i > 0$ is b/k under D, that of the "global rules" $R_{0,j} \to C$ is $b/k + \epsilon$ at all sites and when evaluated globally. As a consequence, under D_i the n rules constructed from R_i are ranked highest by all reasonable utility functions, but globally the rules corresponding to R_0 have a higher utility.

It remains to be shown, that a distribution as described above exists. An additional constraint is that no other rule in $\mathcal{H}$ may reach higher utility scores, neither at any local site nor globally. The following construction is possible if $\mathcal{H}$ contains only single rules $A \to C$ with each A being a conjunction of literals. For k sites and n rules to be selected let

$$z := \lceil \log_2(k+1) \rceil \cdot \lceil \log_2(2n) \rceil.$$

For at least one set of z atomic formulas $\{a_1, \ldots, a_n\}$ it is assumed that

$$\{l_1 \wedge \ldots \wedge l_z \to C \mid l_i \in \{\neg a_i, a_i\} \text{ for } 1 \leq i \leq z\} \subseteq \mathcal{H}.$$

For all considered rules literal l_i refers to the same atomic formula, but it may be positive or negative. Each of the rules may be represented as a boolean vector of length j, where the ith bit refers to the sign of literal i. In turn, each vector v of length j represents a rule $(A_v \to C) \in H$, and for two such vectors $v_i \neq v_j$ it is $A_{v_i} \cap A_{v_j} = \emptyset$.

Now the bit representations can be used to define the sets $R_{i,j}$ for $0 \leq i \leq k$ and $1 \leq j \leq n$ from above: Set the first $\lceil \log_2(k+1) \rceil$ to the binary encoding of the corresponding site number i, and let the subsequent $\lceil \log_2(n) \rceil$ bits encode the rule number j. Each combination of i and j covers two subsets now, since there is one more bit/literal. The subset defined by an even number of

positive literals is defined as positive ($R^+_{i,j}$), the other one as negative ($R^-_{i,j}$). The following equalities imply a common marginal distribution:

$$\int_{x \in R^+_{i,j}} D(x)dx = \int_{x \in R^-_{i,j}} D(x)dx = \frac{1}{(k+1)2n}$$

$$D(x) = D(x') \text{ if } x, x' \in R^+_{i,j} \text{ or } x \in R^+_{i,j} \wedge x' \in R^-_{i,j}.$$
$$D(x) = 0 \text{ if } x \notin \bigcup_{i=0}^{k} \bigcup_{j=1}^{n} \left(R^+_{i,j} \cup R^-_{i,j}\right)$$

For two classes and a default probability of p_0 the joint distribution at site $i \in \{1, \ldots, k\}$ is defined as

$$D_i(x, C) = D(x) \cdot \begin{cases} p_0 + b/k + \epsilon & \text{, for } x \in R^+_{0,j} \\ p_0 - b/k - \epsilon & \text{, for } x \in R^-_{0,j} \\ p_0 + b & \text{, for } x \in R^+_{i,j} \\ p_0 - b & \text{, for } x \in R^-_{i,j} \\ p_0 & \text{, otherwise} \end{cases}$$

for $1 \leq j \leq n$. The positive subsets refer to the original rules, which thus have the desired properties stated earlier[1]. Any rule that covers more than one positive subset will inevitably also cover the negative counterparts. This is a consequence of the syntactical structure of $\mathcal{H}$ and the fact that the bit vectors for positive subsets all have a Hamming-distance of at least two. The Bias will be zero in this case. Specialising rules reduces coverage without any increase in Bias.

The second part of the theorem can be proved similarly. Let the same subsets of $\mathcal{X}$ be associated to $R^+_{0,1} \ldots, R^-_{k,n}$ as before. It is possible to construct a distribution for which all rules have an identical Bias, but which allows to achieve a similar situation as in the first case by locally adjusting the marginal distributions. To this end, let the *local* marginal distributions $D'_i(x)$ for $1 \leq i \leq k$ be defined using the previously defined function $D : \mathcal{X} \to \mathbb{R}^+$, which assigns equal weight to all subsets, and which is uniform within each

[1]If $\log(k+1)$ or $\log(n)$ are no integers, then some subsets of $\mathcal{X}$ are not related to any rule. This has no effect on the validity, since these subsets receive no weight under any of the distributions.

subset:

$$D'_i(x) = D(x) \cdot \begin{cases} 1-\epsilon_m/3 & \text{, for } x \in R^{+/-}_{0,j} & \text{(global rule)} \\ 1 & \text{, for } x \in R^{+/-}_{i,j} & \text{(local rule for site } i\text{)} \\ 1-\epsilon_m & \text{, for } x \in R^{+/-}_{p,j}, p \notin \{0,i\} & \text{(local rule, other site)} \\ 0 & \text{, otherwise} & \text{(unused subset)} \end{cases}$$

with $R^{+/-}_{(\cdot),j} := R^{+}_{(\cdot),j} \cup R^{-}_{(\cdot),j}$. The *local joint* distributions $D'_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ can now be constructed for all sites $1 \leq i \leq k$ using *site-independent* factors:

$$D'_i(x, C) = D'_i(x) \cdot \begin{cases} p_0 + b, \text{ if } x \in R^{+}_{i,j}, 1 \leq j \leq n \\ p_0 - b, \text{ if } x \in R^{-}_{i,j}, 1 \leq j \leq n \\ p_0 \quad \text{, otherwise (BIAS} = 0) \end{cases}$$

All rules have the same BIAS b at all sites, and thus globally. The global COV values are

$$\begin{aligned} \text{COV}_{D'}(R_{0,j} \to C) &= \frac{k(1-\epsilon_m/3)}{k} = 1 - \frac{\epsilon_m}{3} && \text{(global rules)} \\ \text{COV}_{D'}(R_{i,j} \to C) &= \frac{1+(k-1)(1-\epsilon_m)}{k} \leq 1 - \frac{\epsilon_m}{2} && \text{(local rules)} \end{aligned}$$

As required the "global rules" are ranked highest regarding the global distribution D'. At each local site i the corresponding "local rules" $R^{+}_{i,(\cdot)}$ have the highest COV regarding D'_i and are thus ranked highest. More general rules, subsuming several of the positive subsets of $\mathcal{X}$, will have to cover the negative subsets, as discussed in the proof of the first part. Analogously, a specialisation of rules leads to a reduced COV without increasing the BIAS. Choosing ϵ_m so that $\gamma = (1-\epsilon_m)^{-1}$ completes the proof. □

Theorem 1 implies that rules globally performing best are not necessarily among the n locally best rules at *any* site. Even for arbitrarily unskewed data, formalised in terms of definitions 6 and 7, the best rules collected from all sites, including the globally best rules, may be completely disjoint, in the sense that no example is covered twice. Please note that unlike for the case of homogeneously distributed data this is not a problem of misestimation. Theorem 1 applies to arbitrarily large sample sizes.

Although finding the globally best rules from local data is not possible in the worst case, finding approximately best rules might still be tractable.

The following theorem gives a tight bound on the difference between locally and globally evaluated utilities, for simplicity assuming positive utilities and common default probabilities.

Theorem 2 *Let $D : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ denote a global distribution which is a weighted average of k local distributions D_i, all sharing the same default probability of classes. Considering a rule $A \to C \in \mathcal{H}$, let the marginal distributions of D and a local distribution D_i ($i \in \{1, \ldots, k\}$) be factor-similar up to γ for A, and let the joint distributions D and D_i be conditionally similar up to ϵ for the rule. Then the difference between global and local utilities of $\mathrm{Q}^{(\alpha)}$ is bounded by*

$$\begin{aligned}
&\max\left(0, \frac{\mathrm{Q}_{D_i}^{(\alpha)}(A \to C)}{\gamma^\alpha} - \frac{\epsilon}{\gamma^\alpha}\mathrm{Cov}_{D_i}(A \to C)^\alpha\right) \\
\leq &\max\left(0, \mathrm{Q}_D^{(\alpha)}(A \to C)\right) \\
\leq &\max\left(0, \gamma^\alpha \mathrm{Q}_{D_i}^{(\alpha)}(A \to C) + \epsilon\left[\gamma \mathrm{Cov}_{D_i}(A \to C)\right]^\alpha\right)
\end{aligned}$$

For valid choices of ϵ these bounds are tight in the general case.

Proof
A local marginal probability of an antecedent differs by at most a factor of $\gamma^{\pm 1}$ from the corresponding global probability. Similarly, the conditional probability differs by at most an additive constant of $\pm\epsilon$. This implies

$$\begin{aligned}
&\mathrm{Q}_D^{(\alpha)}(A \to C) = \mathrm{Cov}_D(A \to C)^\alpha \mathrm{Bias}_D(A \to C) \\
&\leq \gamma^\alpha \mathrm{Cov}_{D_i}(A \to C)^\alpha \cdot (\mathrm{Bias}_{D_i}(A \to C) + \epsilon) \\
&= \gamma^\alpha \mathrm{Q}_{D_i}^{(\alpha)}(A \to C) + \epsilon\gamma^\alpha \mathrm{Cov}_{D_i}(A \to C)^\alpha
\end{aligned}$$

if all terms are positive. The lower bound is shown analogously.

Given that ϵ is chosen as a valid Bias with respect to the default probability of the target class it is trivial to construct cases for which the bounds are strict. □

For similarly distributed data, e.g. if $\gamma \leq 1.1$ and $\epsilon \leq 0.05$, the bounds are tight enough to allow for estimates with bounded uncertainty. This is illustrated by figures 4-6, showing the bounds for a rule with a global Cov of

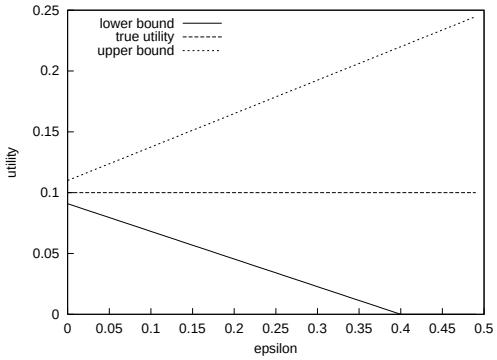

Figure 4: $\mathrm{Q}^{(1)}$ vs. ϵ, $\gamma \leq 1.1$

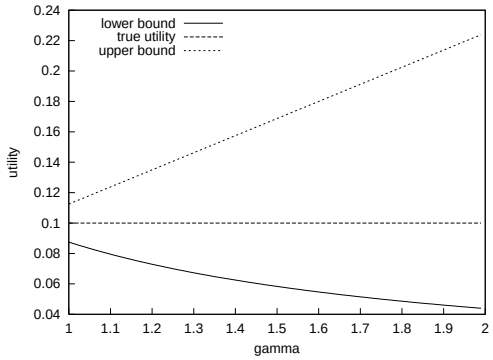

Figure 5: $\mathrm{Q}^{(1)}$ vs. γ, $\epsilon \leq 0.05$

0.25 and a global Bias of 0.4. For $\gamma \leq 1.1$ figure 4 shows upper and lower bounds for $\mathrm{Q}^{(1)}$ with ϵ at the x-axis. Figure 5 and 6 depict bounds for different values of γ, assuming distributions that are conditionally similar up to an $\epsilon \leq 0.05$. Qualitatively the curves for utility function $\mathrm{Q}^{(1)}$ (figure 5) and $\mathrm{Q}^{(1/2)}$ (figure 6) are similar, but the latter is less sensitive to deviating marginal distributions.

Please note that theorem 2 allows to exploit different estimates for each antecedent $A \subset \mathcal{X}$ under consideration. Hence, the theorem is not restricted to learning tasks in which conditional or marginal distributions are known to be very similar. It also allows to collect rule-specific bounds from various sites. Possible sources of rule-dependent bounds on γ and ϵ range from background knowledge over density estimates to previously cached queries.

The question which rules do *not* allow to compute their utilities sufficiently well by techniques related to theorem 2 motivates a new extension of the learning task, discussed in section 5, that explicitly takes the locality of data into account.

5 Relative local subgroup mining

As motivated in the last section, inhomogeneously distributed data allows to define subgroups as subsets of local example sets[2] $\mathcal{E}_i$ that follow different distributions of the target attribute than $\mathcal{E}$ does. This definition of subgroups has a natural interpretation that might be of practical interest in several domains. The corresponding rules could help to point out the char-

[2]More precisely, these definitions refer to the weight of subsets with respect to D and D_i. These weights are of course estimated based on the example sets.

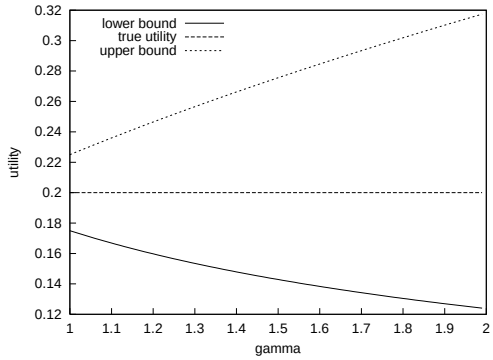

Figure 6: $\mathrm{Q}^{(1/2)}$ vs. γ, $\epsilon \leq 0.05$

acteristics of a single supermarket in contrast to the average supermarket, for example. For the specific case of distributed frequent itemset mining an algorithm for mining *exceptional patterns* taking the locality of data into account has recently been presented [12]. A corresponding extension to the task of rule discovery is lacking. The following function captures the idea of locally deviating rules.

Definition 8 *For $r \in \mathcal{H}$ the utility function $\mathrm{RQ}_{D_i}^{(\alpha)}$ at a site i is defined as*

$$\mathrm{RQ}_{D_i}^{(\alpha)}(r) := \mathrm{Cov}_{D_i}(r)^{\alpha} \cdot (\mathrm{Bias}_{D_i}(r) - \mathrm{Bias}_{D}(r))$$

The rules maximising this function are referred to as relative local subgroups.

Please note that only the global *conditional* distribution is required in this context, since Cov is evaluated locally. Exploiting that Cov differs by at most a factor of γ it is possible to restate theorem 2, again assuming common default probabilities.

Corollary 1 *For a given target class C let*

$$\begin{aligned} rq_{max}^{(\alpha)} &:= \max\{\mathrm{RQ}_{D_i}^{(\alpha)}(r) \mid r \in \mathcal{H},\ r \text{ predicts } C\} \text{ and} \\ rq_{min}^{(\alpha)} &:= \min\{\mathrm{RQ}_{D_i}^{(\alpha)}(r) \mid r \in \mathcal{H},\ r \text{ predicts } C\} \end{aligned}$$

denote the maximal and minimal utilities of relative local subgroups. Then for all rules $r' \in \mathcal{H}$ the difference between local and global utility is bounded by

$$\gamma^{-\alpha} \cdot \left(\mathrm{Q}_{D_i}^{(\alpha)}(r') - rq_{max}\right) \leq \mathrm{Q}_{D}^{(\alpha)}(r') \leq \gamma^{\alpha} \cdot \left(\mathrm{Q}_{D_i}^{(\alpha)}(r') - rq_{min}\right)$$

if all terms are positive.

Corollary 1 allows to translate the utility of local subgroups into global scores with bounded uncertainty for any rule-dependent γ. The special case of a common marginal distribution is obtained by setting $\gamma = 1$.

Corollary 2 *For $\gamma = 1$ the three utility functions for local, relative local, and global subgroup discovery complete each other:*

$$\mathrm{Q}_D^{(\alpha)}(A \to C) = \mathrm{Q}_{D_i}^{(\alpha)}(A \to C) - \mathrm{RQ}_{D_i}^{(\alpha)}(A \to C)$$

Obviously, the tasks of discovering relative local subgroups and that of approximating the global conditional distribution are of similar complexity in this case. Corollary 2 shows how to detect global subgroups searching locally, given precise estimates of $\mathrm{RQ}_{D_i}^{(\alpha)}$, and how to compute $\mathrm{RQ}_{D_i}^{(\alpha)}$ from $\mathrm{Q}_D^{(\alpha)}$ for $\gamma = 1$.

6 Practical considerations

This section relates the subtasks to known learning strategies. One can distinguish between three kinds of strategies, applying trained models, searching exhaustively, and sampling with respect to the global distribution. After discussing these issues it is exemplarily illustrated in this section, how theorem 2 allows to translate local utilities into global ones.

6.1 Model-based search

The idea of a model-based search is to first train a model that approximates the global conditional distribution of the target attribute. If the model yields precise estimates, then $\mathrm{RQ}_{D_i}^{(\alpha)}$ (Def. 8) can directly be computed from the local data, which allows to discover the relative local subgroups in the next step. For a common marginal distribution of $\mathcal{X}$ ($\gamma = 1$) this also allows to discover the global subgroups by applying corollary 2. In the general case bounded estimates for global rule utilities can be given (Cor. 1).

A simple learner that allows to approximate the conditional distribution is Naïve Bayes. It can easily be applied to distributed data, because the global model can be obtained by collecting the counts from all sites. A more complex technique that usually comes with higher accuracy is distributed boosting. An algorithm similar to confidence-rated versions of AdaBoost [8]

has been presented in [7]. Please note that skewed data is a known source of failure in the scope of distributed boosting.

Another problem with the model-based strategy is that even if the model is precise, it can hardly be expected to reach 100% accuracy in practice. This means that some of the relative local subgroups may not be found, since it is unknown for which subsets the predictions of the model are poor. Hence, it is a heuristic rather than a probabilistic search strategy.

6.2 Searching exhaustively

The fact that an approximation of the conditional distribution does not help to find global subgroups reliably in the general case justifies to address relative local and global subgroup discovery by exhaustively searching the hypothesis space.

For frequent itemset mining efficient distributed strategies exist [11], basically exchanging itemsets and counts. Some of the pruning strategies allow to generate candidates for relative local subgroups, since the pruning based on counts received from other sites affects itemsets that are locally more frequent than globally. Local and global subgroups are finally obtained by applying the utility function to the results. The disadvantage is that there will usually be many more frequent itemsets than subgroups, because the pruning performed during itemset mining does not take into account the specific choice of a utility function.

Applying the pruning strategy of MIDOS [10] allows to safely discard specialisations of a rule with small Cov, if these cannot contain improvements on the best n subgroups found so far. Additionally, since global counts generally need to be collected from all sites, more specific pruning techniques sometimes allow to stop the evaluation of a rule after receiving the counts of some of the sites, already.

6.3 Sampling from the global distribution

As discussed in the introduction it is often not possible to collect all the data at a single site. If the reasons are communication costs rather than privacy, then it may still be cheaper to learn directly with respect to the global distribution than to address a hard learning task using distributed approaches that do not come with any guarantees. Applying the adaptive sampling techniques proposed in [9], one can hope that probabilistic guarantees can

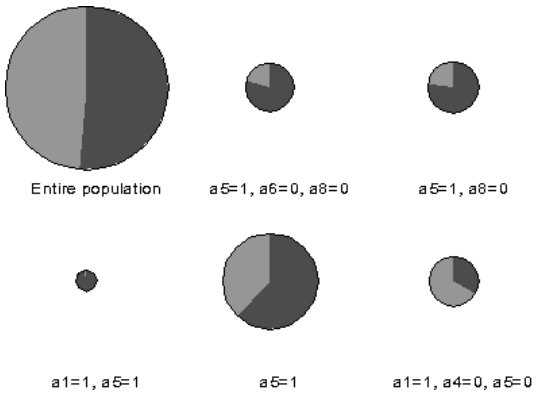

Figure 7: Visualisation of 5 best subgroups

be given after transferring just a small fraction of the data to a central node for the data mining step.

6.4 Estimating Utilities with Bounded Uncertainty

This subsection describes a first experiment that illustrates some of the presented ideas. Due to a lack of publicly available datasets for distributed Data Mining, synthetic data was used. As an advantage, this allows to control the different kinds of skews.

To prepare the data, a decision tree for a domain of 10 boolean attributes has been constructed at random. For each inner node the probability of the tested attribute being 1 was fixed to a value randomly drawn from $N(0.5, 0.25)$. The same was done for the distribution of the boolean target label at the leaves. For all examples unspecified attributes were simply completed by drawing truth values uniformly. The examples were distributed to 5 sites by explicitly assigning a separate γ- and ϵ-skew to each leaf for each site. The skew-parameters were selected uniformly within the previously used intervals: $\gamma \in [0, 1.1]$, $\epsilon \in [0, 0.05]$. Based on this randomly constructed tree 10.000 examples were generated as an input to the following subgroup discovery experiments.

The MIDOS algorithm, part of the KEPLER toolbox, was applied to the data, in order to select 5 best subgroups according to $\mathrm{Q}^{(1)}$. In figure 7 these subgroups are visualised by circles. The colours reflect the conditional distributions of the target, while the sizes represent coverage. Each dot in

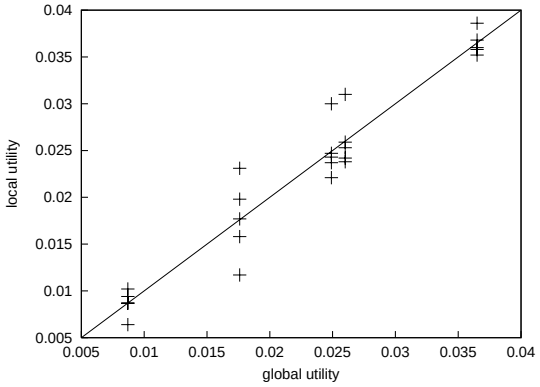

Figure 8: Global vs. local evaluation.

figure 8 compares the global utility of a rule (x-axis) to the corresponding local utilities at all sites (y-axis). Dots close to the diagonal represent similar utilities, which are useful for estimating the global utilities from local ones with bounded uncertainty. Table 1 lists the bounds that could be derived based on the local estimates, exploiting $\gamma \leq 1.1$ and $\epsilon \leq 0.05$. It is interesting to note, that only for the largest subgroup ($a5 = 1$, Cov ≈ 0.35) the bounds are useless, because for large subgroups the utility can easily be estimated from samples, instead. In contrast, the smallest of the 5 selected subgroups has a Cov of below 2%. This illustrates why generating all frequent itemsets is often an inefficient approach to subgroup discovery.

Subgroup	**global** $\mathrm{Q}^{(1)}$	**Lower b.**	**Upper b.**
a5=1, a6=0, a8=0	0.0249	0.0231	0.0292
a5=1, a8=0	0.0260	0.0235	0.0316
a1=1, a5=1	0.0087	0.0083	0.0105
a5=1	0.0365	0.0191	0.0573
a1=1, a4=0, a5=0	0.0176	0.0164	0.0180

Table 1: Bounds due to theorem 2.

7 Conclusion

The behaviour of different rule selection metrics, their similarity for various skews and how well they may be estimated from samples has been investigated in the recent years. What is lacking is an investigation of how these metrics behave in the scope of distributed learning. This paper is a first step into this direction. First of all it has been shown that the utility measures common in the literature on subgroup discovery can be applied to homogeneously distributed data in the same way as to a single example set. If the different sites do not share a single underlying distribution generating the data, however, then even precise estimates may yield completely disjoint rulesets at all sites, none of which contains a single one of the best n rules. For the general case a tight bound for the difference between global and local rule utilities has been proven, which allows to translate local rule utilities into global ones with bounded uncertainty. For the task of discovering rules that have a higher local than global utility it has been shown that it is at least as hard as approximating the global conditional distribution of the target attribute. For a common marginal distribution one problem can be solved locally, given a solution for the other one.

The results indicate that distributed subgroup discovery is a hard problem, since it requires precise estimates of both, the global marginal and the global conditional distribution. The former may e.g. be obtained by distributed variants of frequent itemset mining, the latter by means of distributed boosting. As discussed there are good reasons, however, to tackle the problem by exhaustively searching the hypothesis space, applying specific pruning strategies wherever possible.

Future work will compare concrete implementations empirically, using synthetic and real-world data in combinations with different ways to distribute examples.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft, Collaborative Research Centre 475 on *Reduction of Complexity for Multivariate Data Structures*.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large data bases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, pages 478–499, Santiago, Chile, sep 1994.

[2] D. W.-L. Cheung and Y. Xiao. Effect of Data Skewness in Parallel Mining of Association Rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 48–60, 1998.

[3] P. A. Flach. The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. Morgen Kaufman, 2003.

[4] J. Fürnkranz and P. Flach. ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms. *Machine Learning*, 58(1):39–77, 2005.

[5] J. Fürnkranz and P. A. Flach. An Analysis of Rule Evaluation Metrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. Morgen Kaufman, 2003.

[6] W. Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 3, pages 249–272. AAAI Press/The MIT Press, Menlo Park, California, 1996.

[7] A. Lazarevic and Z. Obradovic. Boosting algorithms for parallel and distributed learning. *Distributed and Parallel Databases Journal*, 11(2):203–229, 2002.

[8] R. E. Schapire and Y. Singer. Improved Boosting Using Confidence-rated Predictions. *Machine Learning*, 37(3):297–336, 1999.

[9] T. Scheffer and S. Wrobel. Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling. *Journal of Machine Learning Research*, 3:833–862, 2002.

[10] S. Wrobel. An Algorithm for Multi–relational Discovery of Subgroups. In J. Komorowski and J. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery: First European Symposium (PKDD 97)*, pages 78–87, Berlin, New York, 1997. Springer.

[11] M. J. Zaki. Parallel and Distributed Association Mining: A Survey. *IEEE Concurrency*, 7(4):14–25, /1999.

[12] S. Zhang, C. Zhang, and J. Yu. An Efficient Strategy for Mining Exceptions in Multi-databases. *Information Sciences*, 1-2(165):1–20, 2004.